5.6 High-Speed Interconnect for a Multiprocessor Server Using Over 1Tb/s Crossbar

Jun Yamada, Hiroyuki Adachi, Yutaka Mori, Akihiko Harada, Seishi Okada, Hisashige Ando

Fujitsu, Kawasaki, Japan

A high-bandwidth inter-processor interconnect is developed for Fujitsu's PRIMEQUEST server [1] with a high-speed low-power small-die-area backplane interconnect technology named Mori/Muta transceiver logic (MTL). Using MTL, a crossbar LSI with a total I/O bandwidth exceeding 1Tb/s is realized.

Figure 5.6.1 shows a high-level block diagram of the server system. The backplane crossbar consists of two components, an address crossbar and a data crossbar. The address crossbar consists of 2 global address crossbar LSIs (GACs). The data crossbar consists of 4 sets of global data crossbar LSIs (GDX pairs). These pair of GACs and/or GDXs can operate in duplicated mirror mode for high-reliability operation or operate in non-mirror mode to achieve a maximum system bandwidth of 170GB/s. The specifications of both GAC and GDX LSIs are listed in Fig. 5.6.2 (a).

The data-transfer speed between these crossbars and the system board is 1.333Gb/s and the total I/O pin bandwidths of the LSIs are 1.26Tb/s for GAC and 0.99Tb/s for GDX.

In order to satisfy the high bandwidth requirement with a small number of LSIs, it is important to have a small and low-power data transmission driver and receiver. The wiring density and the ease of PCB wiring is another important requirement as there are many other digital signal wirings on the PCB connecting other components. MTL is designed to address the above requirements. The MTL specifications are shown in Fig. 5.6.2 (b). MTL uses a low-swing single-ended 1.333Gb/s transmission with a 3-tap pre-emphasis for the driver. It is source synchronous with one differential clock connection for each 36 data signals.

Figure 5.6.3 shows a comparison between the MTL and the previously published high-speed data-transmission technologies [2, 3]. Although the wire speed of the MTL is not fast compared to the other high-speed differential transmission technologies, but MTL has impressive mW/Gb/s and mm^2/Gb/s metrics. Slower per-wire bit-rate of the MTL reduces the skin-effect high-frequency loss and allows to use a narrow 130µm-wide PCB trace for 60cm signal transmission and a 70µm-wide trace for shorter distance. The slower data rate also makes the synchronization of a wide parallel data-bus implementation easier.

The schematic of the MTL driver and receiver are shown in Fig. 5.6.4. The receiver has an on-die termination. The output from the 3 stages of the differential amplifier drives both a data sampling latch and a clock input of a phase detector. The phase detector samples the data sampling clock as shown in Fig. 5.6.5. Since the clock is sampled at the data edge, this scheme is called data synchronous. When the clock edge is too early compared to the center of the data eye, the clock is sampled "low". When the clock is too late, the clock is sampled "high". The feedback loop that consists of a 1/30 divider and a digital delay line generates the data sampling clock by delaying the source-synchronous clock input. The delay of the digital delay line increases when the "low" sample continues and the clock edge of each data sampling clock is centered to the data eye. The same feedback loop advances the data sampling clock when the "high" sample continues. Since the clock is sampled at every data edge and the phase error is fed back for compensation, the adjustment is continuous. The divider counts down the phase detector output to optimize the loop response. The resolution of the 5b digital delay line is 30ps (fast PVT corner) and the adjustable range is about 1.0ns. As this adjustment is done for each data input, the design tolerates more than 10cm of wire-length difference among the 36b group sharing the same clock.

The length of the wires the MTL drives in the system is between 30cm and 60cm with at most two connectors. The MTL driver is optimized to operate for this range of transmission-line length and the fixed level pre-emphasis is designed to give an optimal waveform for the longest 60cm transmission. The eye diagram for 60cm transmission is shown in Fig. 5.6.6 (a) with pre-emphasis and (b) without pre-emphasis. As shown in Fig. 5.6.6 (c), this fixed level pre-emphasis design slightly over-compensates the 30cm transmission, but this is acceptable. This fixed level pre-emphasis design helps to reduce the die area of the driver. Also the received waveform is good enough for the use of the data-synchronous phase detector clock with the above mentioned 30 to 60cm transmission. Figure 5.6.6 (d) shows the input data, adjusted sampling clock, and output data waveforms. The sampling clock traces the delay variations of the input data and centers the data sampling point with the data-synchronous scheme.

It is important to have a good return current path for the single-ended transmission to be successful. All the MTL traces in the PCB and the package are sandwiched between the ground and V_{dd} planes. There are enough C4 bumps for V_{dd} and ground connections. The area of the MTL driver and receiver macros are adjusted to occupy two C4 bump areas. A large percentage of both macro areas is occupied by decoupling capacitors. The 100pF decoupling capacitor is used for each MTL driver and the 60pF decoupling capacitor is used for each receiver for bypassing between V_{dd} and ground.

The die micrographs of the GAC and the MTL driver/receiver layout are shown in Fig. 5.6.7.

By using the small low-power MTL driver and receiver, 704 drivers and 352 receivers, consuming 11.6W with total I/O bandwidth exceeding 1Tb/s, are successfully integrated on the GAC LSI. The 32 CPU socket multiprocessor crossbar with 170GB/s bandwidth is realized with the total of 2 GAC and 8 GDX LSIs.

Acknowledgements:
The authors would like to acknowledge A. Kabemoto and T. Noda who headed PRIMEQUEST server development for their support of MTL development. The authors would like to thank many engineers who have involved in the development in Fujitsu. Special thanks to T.Muta who developed the logic circuit around the MTL for high-level data alignment. A part of this work was funded by the New Energy and Industrial Technology Organization of Japan.

References:
[1]T. Shimizu, et al., "Fujitsu PRIMEQUEST: 32way SMP Open Servers with Powerful Reliability Features," *The International Conference on Dependable Systems and Networks*, supplemental volume, pp.104-111, June, 2005.
[2]K. Yamaguchi, et al., "12Gb/s Duobinary Signaling with x2 Oversampled Edge Equalization," *ISSCC Dig. Tech. Papers*, pp. 70-71, Feb., 2005.
[3]R. Payne, et al., "A 6.25Gb/s Binary Adaptive DFE with First Post-Cursor Tap Cancellation for Serial Backplane Communication," *ISSCC Dig. Tech. Papers*, pp. 68-69, Feb., 2005.

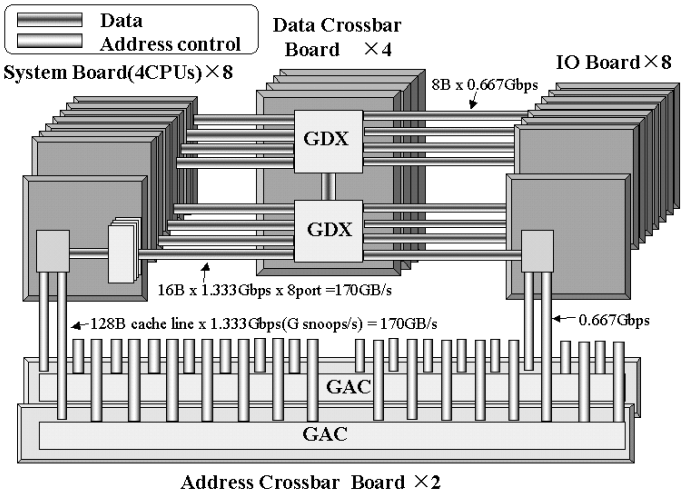

Figure 5.6.1: System configuration.

		GAC	GDX
Process Technology		90nm CMOS	
Chip Size(mm^2)		13.5x15.0	12.5x12.5
Power Supply		1.2V/2.5V	
Power Consumption		25.7W(IO=11.6W)	18.7W(IO=9.3W)
Package Technology		C4/BGA	
Pins	MTL (1.33G/667M/CLK)	1120 (832 / 224 / 64)	888 (640 / 200 / 48)
	Others signals	77	71
	Power/Gnd	919	722
MTL I/O Band Width (Tbps)		1.26	0.99

(a) Crossbar LSI Specifications

		Spec.
Physical Design	IO Macro Size	176.4 × 352.8μm^2
	Power Consumption	25mW@1.3Gbps
Transmission	Data Rate	1.333Gbps
	Connection	Single-end, 1:1
	Driver	CMOS with Pre-Emphasis
	Receiver, Swing	HSTL, 700mV

(b) MTL Specifications

Figure 5.6.2: Crossbar LSI and MTL specifications.

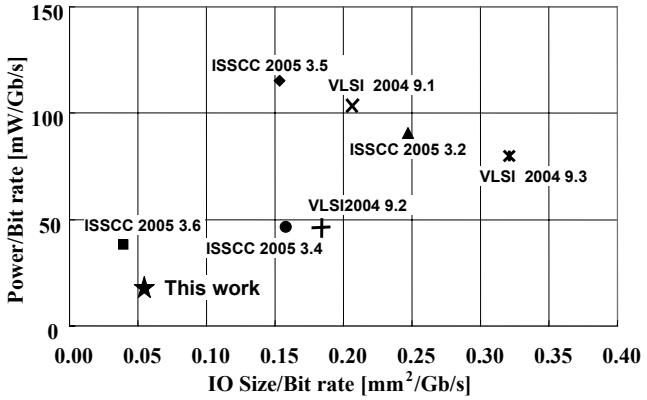

Figure 5.6.3: Comparison between MTL and previous works.

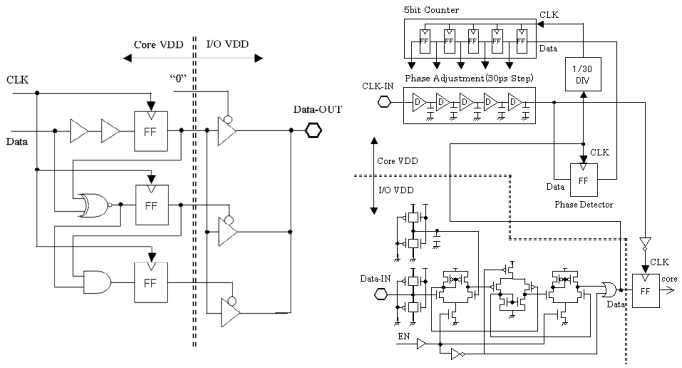

Figure 5.6.4: MTL driver and receiver circuits.

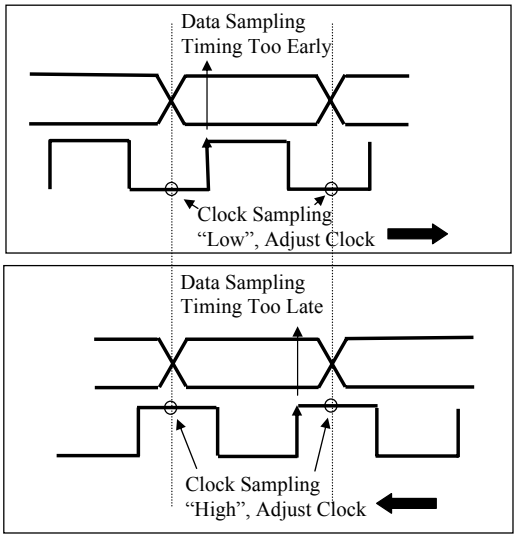

Figure 5.6.5: Data-synchronous scheme.

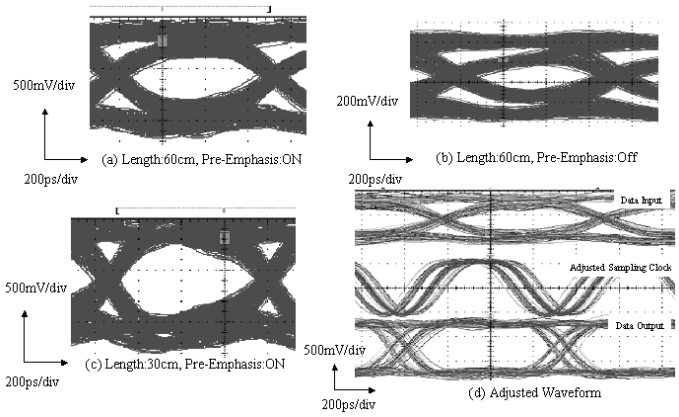

Figure 5.6.6: Eye diagram and adjusted waveform.

Continued on Page 641

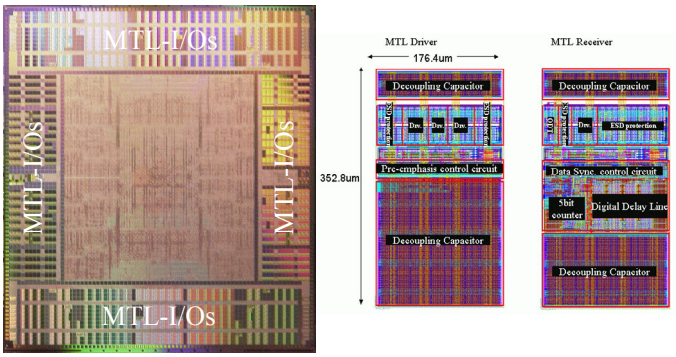

Figure 5.6.7: GAC die micrograph and MTL driver/receiver layout.

 1-4244-0079-1/06/$20.00